Neural Text ~~de~~Generation

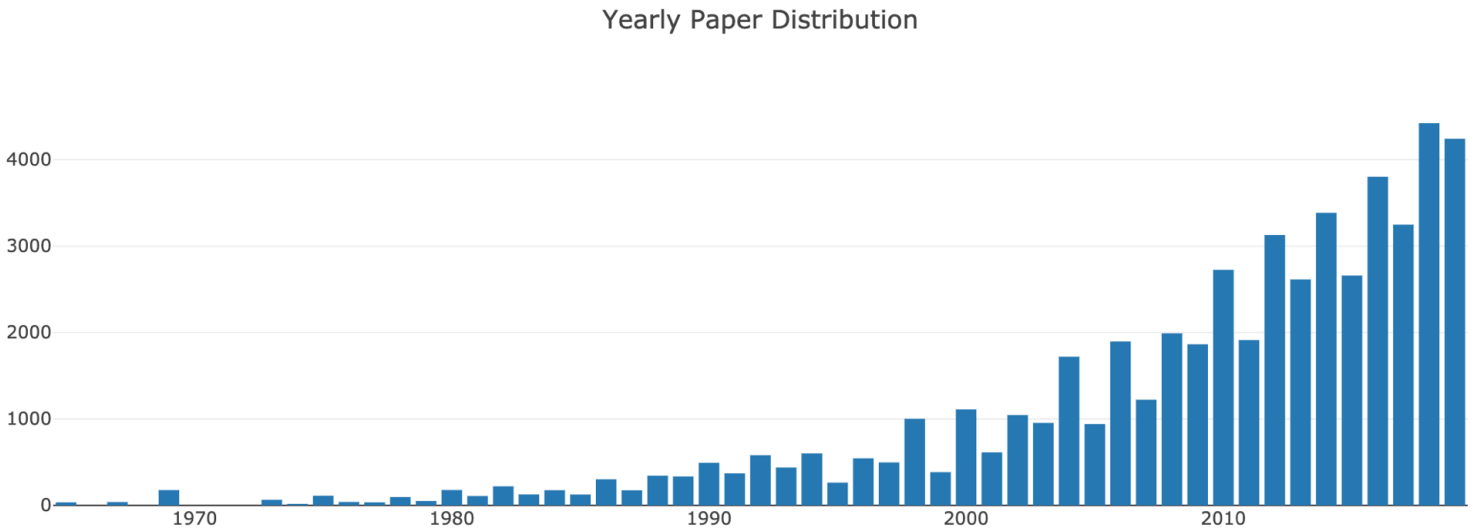
Inconsistency & Unlikelihood

Sean Welleck

NYU

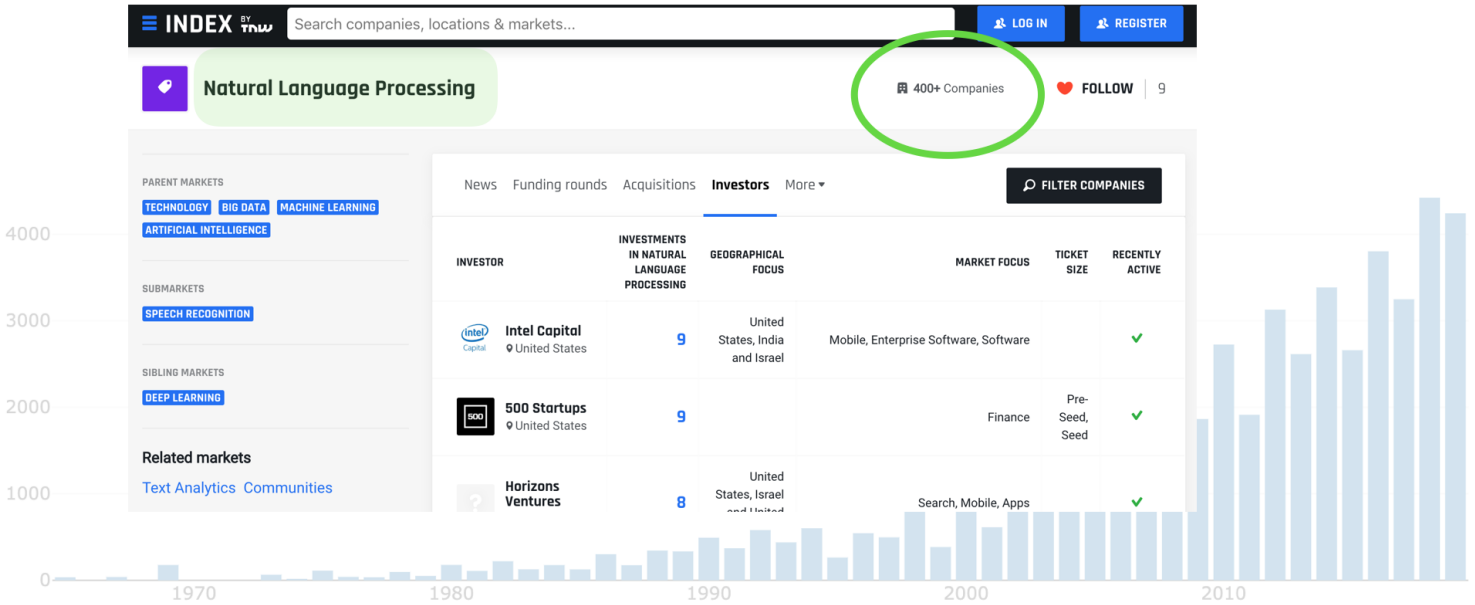
18th March, 2020

- Despite substantial progress...



nlpexplorer.org

- Despite substantial progress...



- Despite substantial progress...

Better Language Models and Their Implications

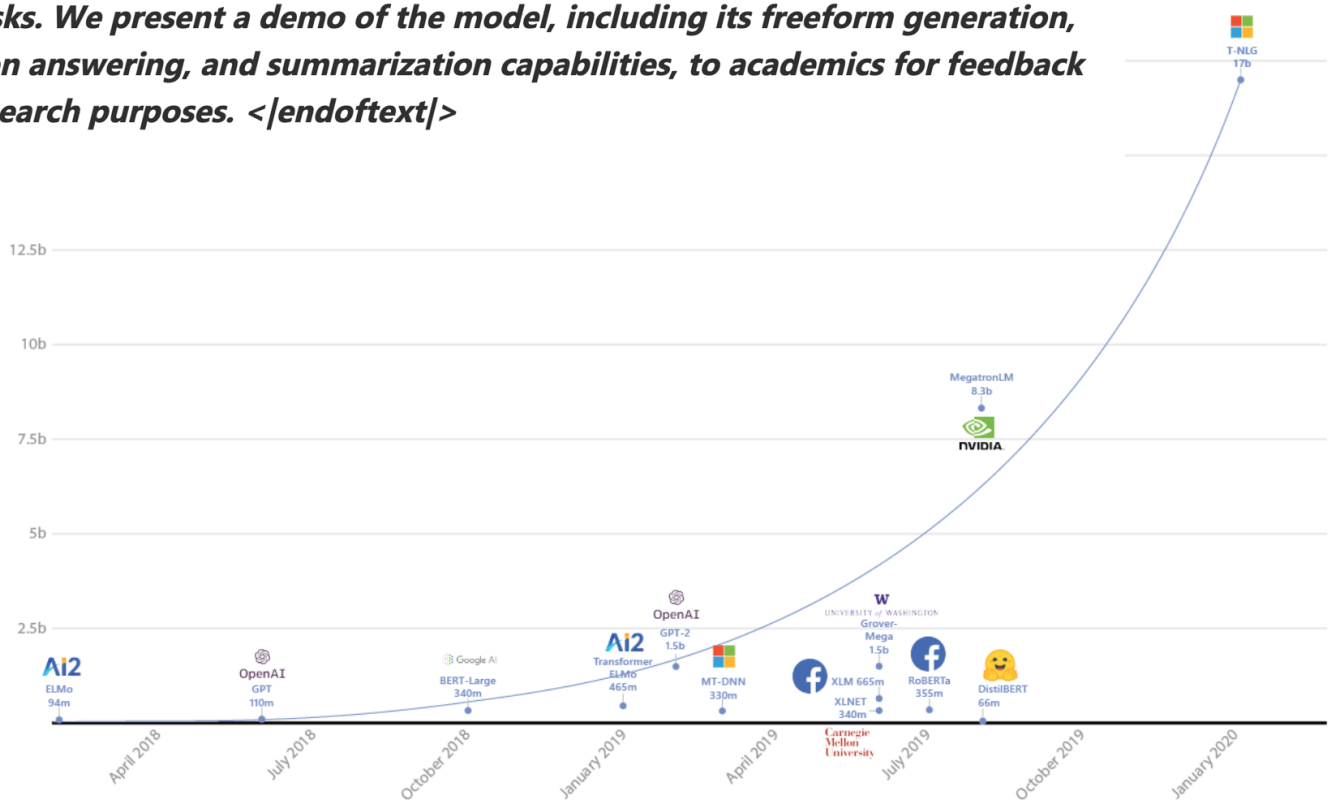
We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter: "we

https://openai.com/blog/better-language-models/

- Despite substantial progress...

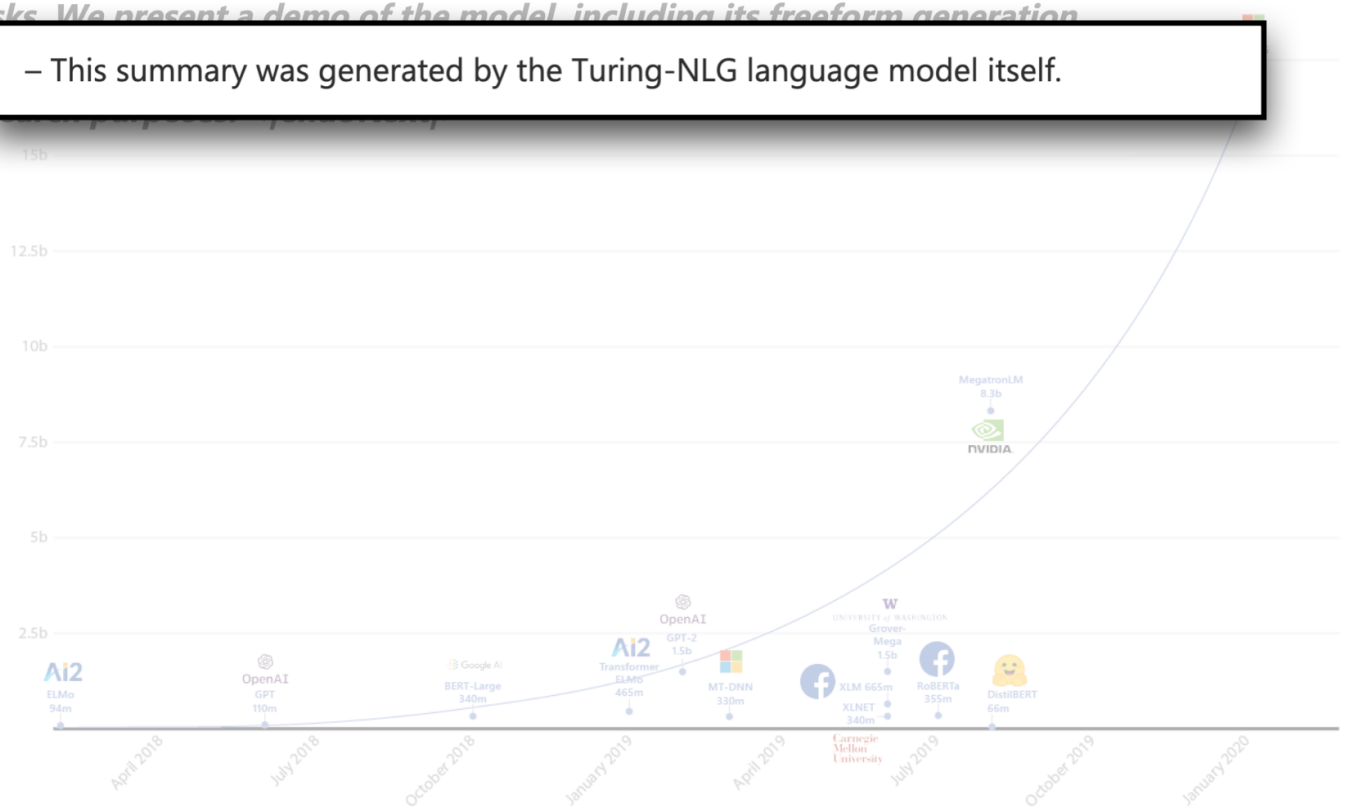
Turing Natural Language Generation (T-NLG) is a 17 billion parameter language model by Microsoft that outperforms the state of the art on many downstream NLP tasks. We present a demo of the model, including its freeform generation, question answering, and summarization capabilities, to academics for feedback and research purposes. <|endoftext|>



- Despite substantial progress...

Turing Natural Language Generation (T-NLG) is a 17 billion parameter language model by Microsoft that outperforms the state of the art on many downstream NLP tasks. We present a demo of the model, including its freeform generation

– This summary was generated by the Turing-NLG language model itself.



- Standard neural language models show degeneration

Context:

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Continuation (BeamSearch, b=10):

"The unicorns were able to communicate with each other, they said unicorns. a statement that the unicorns. Professor of the Department of Los Angeles, the most important place the world to be recognition of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the…

GPT-2 Continuations
[from Holtzmann et al ICLR 2020]

- Standard neural language models show degeneration

Prefix
$\mathcal{L}_{\text{MLE}}$

... starboard engines and was going to crash . " We 're going in ,"
he said . " We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to

Repetition
Model: 40% 4-gram repetition
Human: ~1% 4-gram repetition

Large-Scale Transformer LM [from Welleck et al ICLR 2020]

This talk

- What do we mean by 'neural text degeneration'? (**Background**)
- What do we theoretically know about it? (**Inconsistency**)
- What are some ways of preventing it? (**Inconsistency, Unlikelihood**)

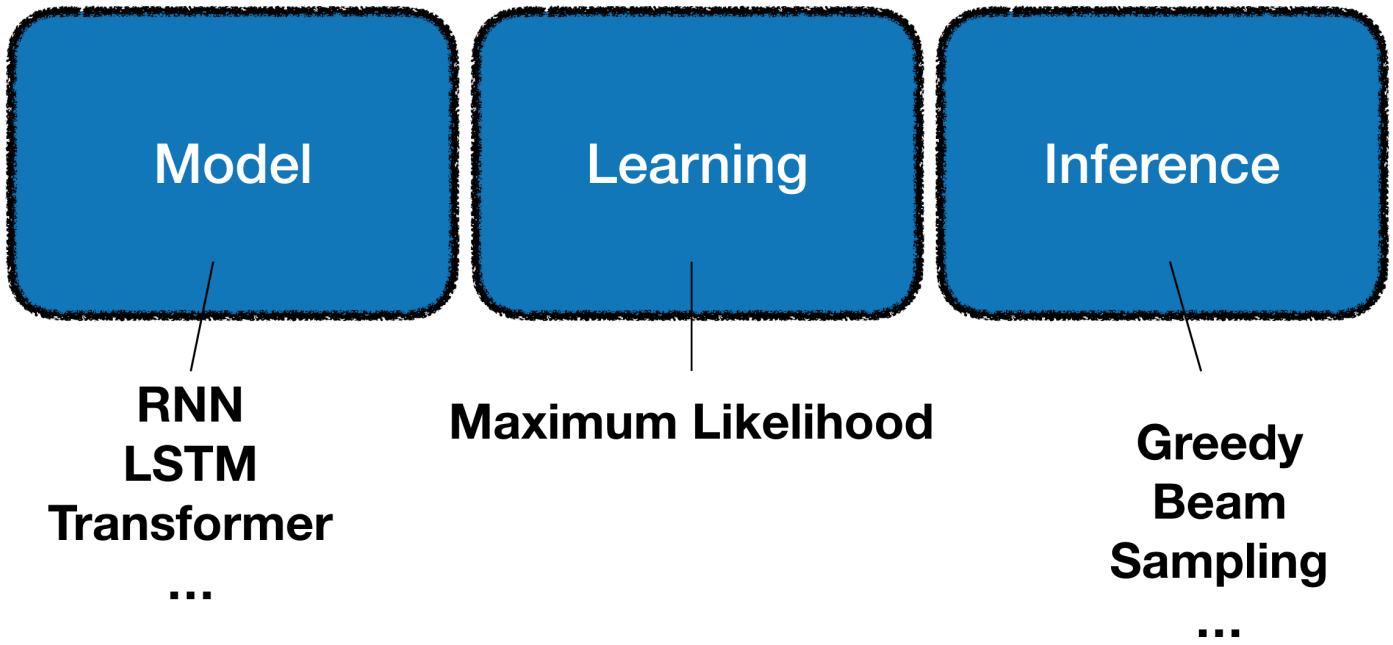
Background – Neural Sequence Modeling

Goal:

1. estimate $p_\theta(\underbrace{y_1, \ldots, y_T}_{\text{sequence}} | \underbrace{C}_{\text{context}})$
   - $p_\theta(\texttt{how, are, you, ?}|\text{元気ですか?})$
2. decode $(\underbrace{\hat{y}_1, \ldots, \hat{y}_T}_{\text{sequence}}) \sim \underbrace{\mathcal{F}(p_\theta, C)}_{\text{decoding algorithm}}$
   - $(\texttt{how, are, you, ?}) \sim \mathcal{F}(p_\theta, \text{元気ですか?})$

Machine translation, dialogue modeling, language modeling, story generation, text completion, etc.

Background – Neural Sequence Modeling

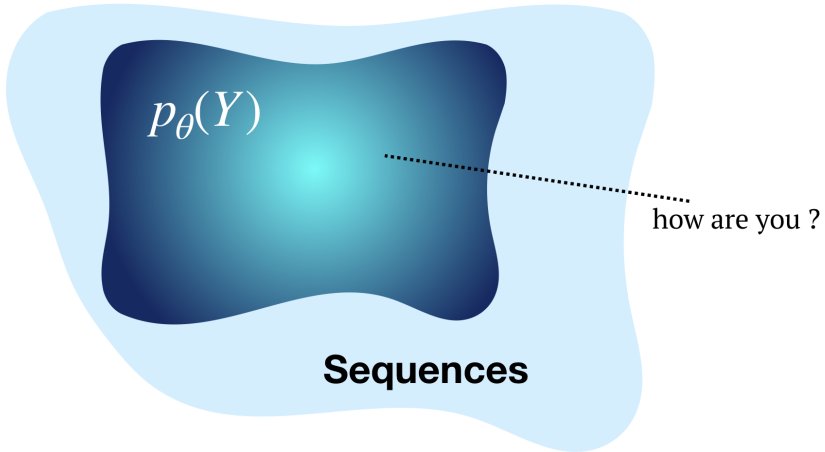


Model

- Neural autoregressive sequence model:

$$p_\theta(y_1, \ldots, y_T | C) = \prod_{t=1}^{T} \underbrace{p_\theta(y_t | y_{<t}, C)}_{\text{neural network}}$$

- Neural network:
  - RNN / LSTM
  - Transformer



Learning

Maximum Likelihood Estimation (MLE)

1. Given dataset $\{(C^{(n)}, Y^{(n)})\}_{n=1}^{N}$,

$$\arg\max_{\theta} \sum_{n=1}^{N} \sum_{t=1}^{T^{(n)}} \log p_\theta(y_t^{(n)} | y_{<t}^{(n)}, C^{(n)})$$

Inference

1. Given trained model p_θ, context C

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} \log p_\theta(Y|C)$$

Inference

1. Given trained model p_θ, context C

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} \log p_\theta(Y|C)$$

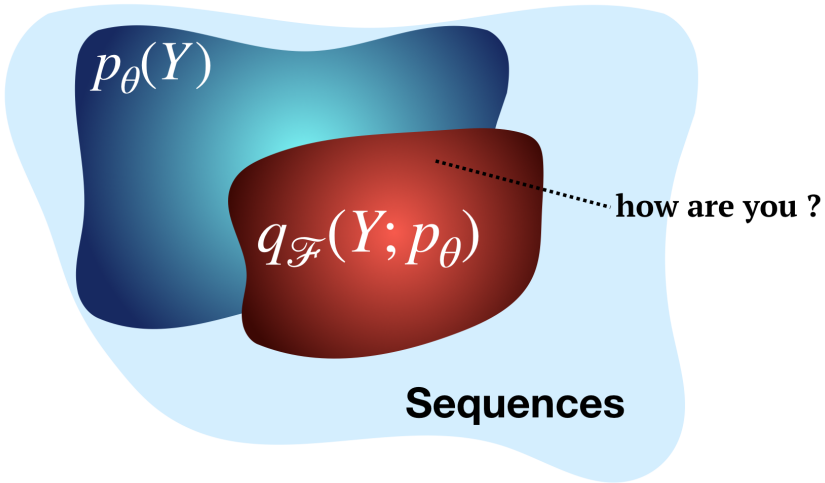
2. **Approximate** using a decoding algorithm $\mathcal{F}(p_\theta, C)$
   - Greedy search
   - Beam search
   - Sampling

Inference

1. Given trained model p_θ, context C

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} \log p_\theta(Y|C)$$

2. **Approximate** using a decoding algorithm $\mathcal{F}(p_\theta, C)$
   - Greedy search
   - Beam search
   - Sampling
3. The decoding algorithm induces a new sequence distribution



Inference – Decoding Algorithms

Greedy Search

Recursively selects the most likely token until $\langle \text{eos} \rangle$:

$$\tilde{y}_t = \arg\max_{v \in V} \log p_\theta(y_t = v \mid \tilde{y}_{<t}, C).$$

Inference – Decoding Algorithms

Greedy Search

Recursively selects the most likely token until $\langle \text{eos} \rangle$:

$$\tilde{y}_t = \arg\max_{v \in V} \log p_\theta(y_t = v \mid \tilde{y}_{<t}, C).$$

Beam Search

Searches with a set of k high scoring prefixes:

$$\mathrm{P}_t = \bigcup_{\rho \in \mathrm{P}_{t-1}^{\text{top}}} \{\rho \circ v \mid v \in V\}$$

$$\mathrm{P}_t^{\text{top}} = \arg\text{top-k}_{\rho \in \mathrm{P}_t} s(\rho).$$

Inference – Decoding Algorithms

Ancestral Sampling

$$\tilde{y}_t \sim p_\theta(y_t \mid \tilde{y}_{<t}, C).$$

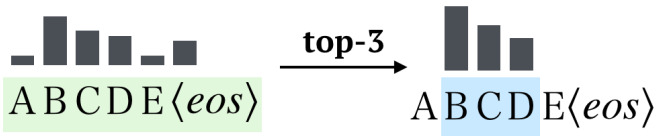
Inference – Decoding Algorithms

Ancestral Sampling

$$\tilde{y}_t \sim p_\theta(y_t \mid \tilde{y}_{<t}, C).$$

Top-k Sampling

$$q(v) \propto \begin{cases} p_\theta(v \mid y_{<t}, C), & v \in \text{arg top-k } p_\theta, \\ 0, & \text{otherwise.} \end{cases}$$



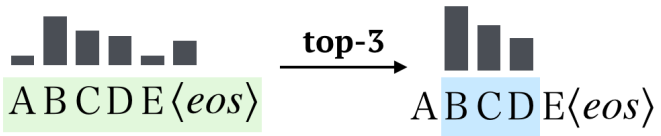
Inference – Decoding Algorithms

Ancestral Sampling

$$\tilde{y}_t \sim p_\theta(y_t \mid \tilde{y}_{<t}, C).$$

Top-k Sampling

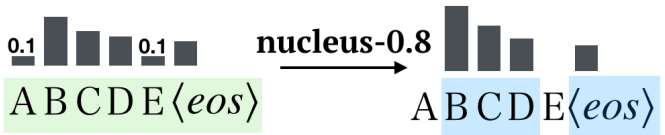
$$q(v) \propto \begin{cases} p_\theta(v \mid y_{<t}, C), & v \in \text{arg top-k } p_\theta, \\ 0, & \text{otherwise.} \end{cases}$$



Top-p ('nucleus') Sampling

$$q(v) \propto \begin{cases} p_\theta(v \mid y_{<t}, C), & v \in V_\mu, \\ 0, & \text{otherwise.} \end{cases}$$

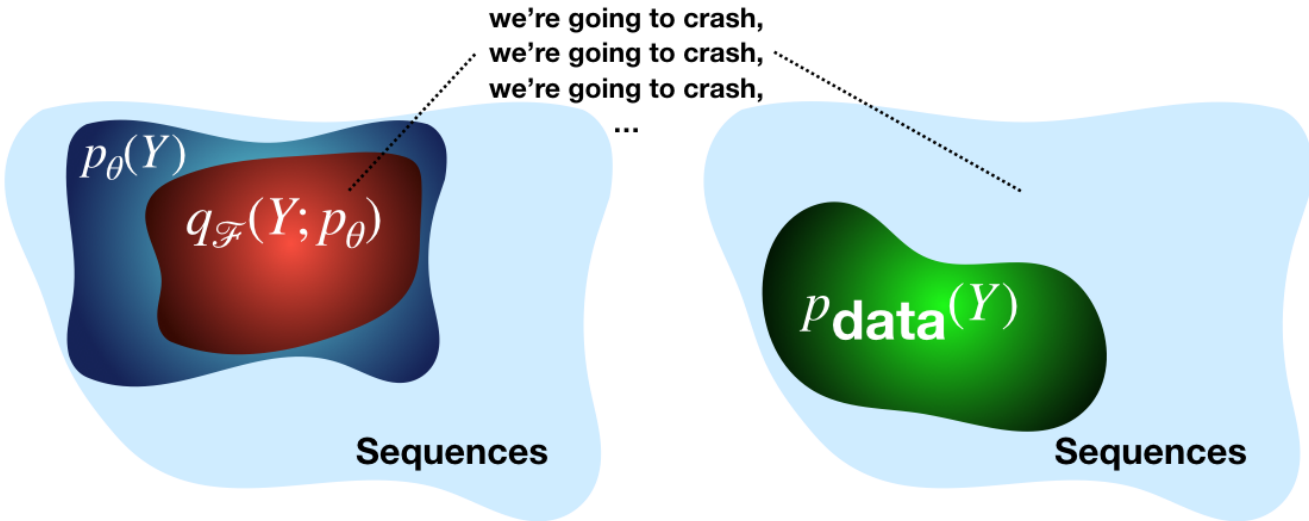
V_μ : smallest set with total prob $< \mu$



Problem: 'Text Degeneration'

- Language models trained with **maximum likelihood** produce text which **differs from the true distribution** when using **common decoding algorithms**:

$$q_{\mathcal{F}}(Y) \neq p_{\text{data}}(Y)$$



Symptoms

- **Repetition**

Prefix	*... starboard engines and was going to crash . " We 're going in ,"*
$\mathcal{L}_{\text{MLE}}$	he said . " We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to

Repetition
Model: 40% 4-gram repetition
Human: ~1% 4-gram repetition

Large-Scale Transformer LM [from Welleck et al ICLR 2020]

Symptoms

- **Repetition**
- **Non-terminating sequences**

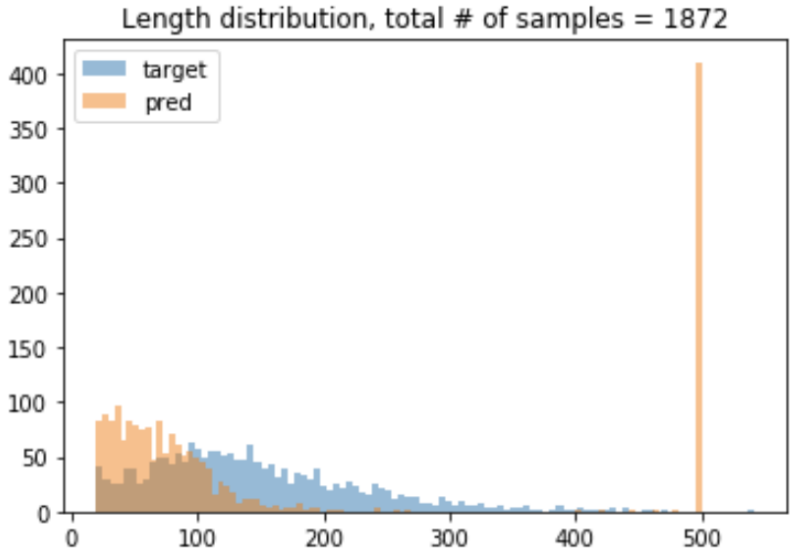


Figure: GPT-2 Completions (Wikitext-103 fine-tuned)

Symptoms

- **Repetition**
- **Non-terminating sequences**
- **Logical contradictions**

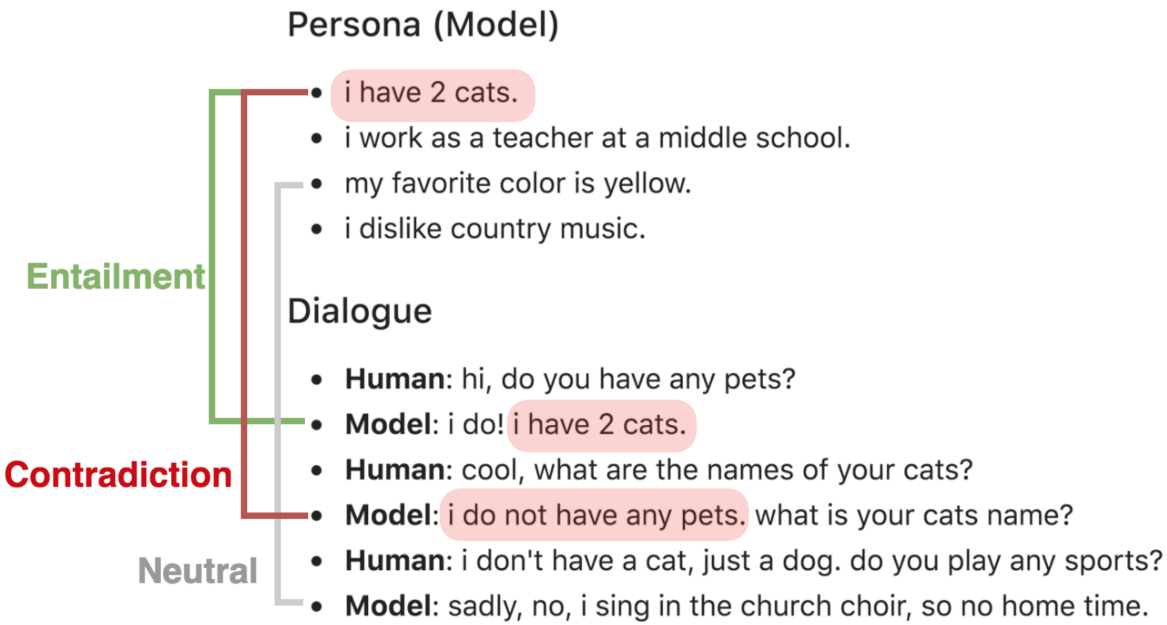


Figure: Welleck et al 2019, **Dialogue NLI**

Symptoms

- **Repetition**
- **Non-terminating sequences**
- **Logical contradictions**

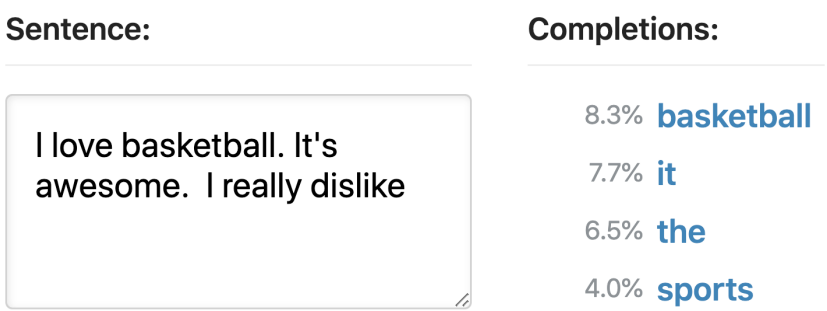
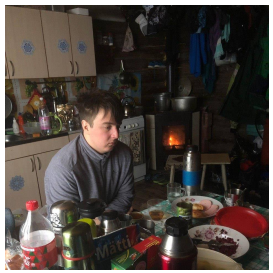
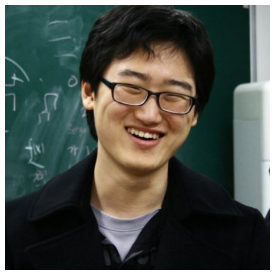


Figure: GPT-2 completion

Consistency of a Recurrent Language Model with Respect to Incomplete Decoding

Welleck, Kulikov, Kim, Pang, Cho 2020 (under review)

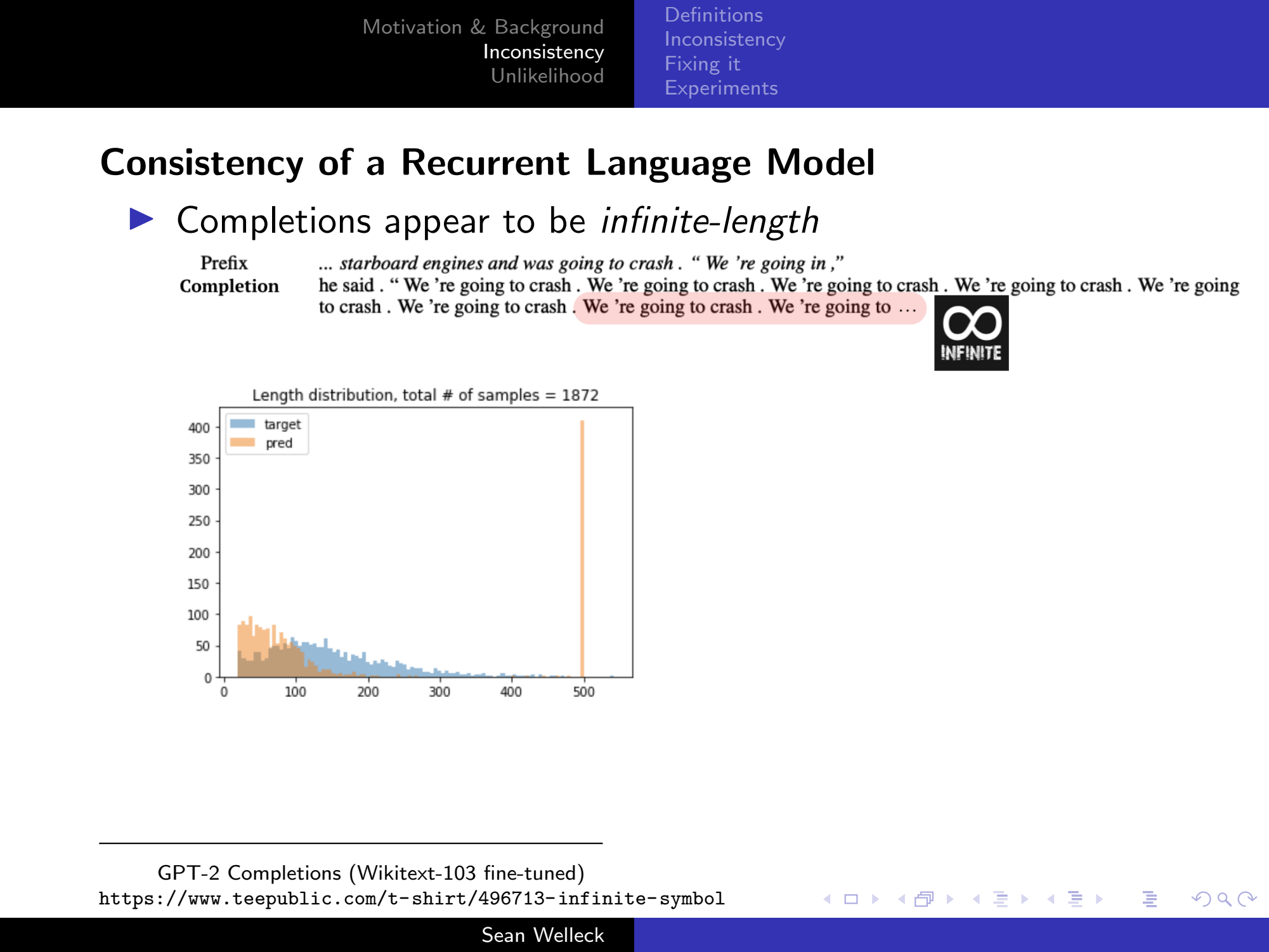
https://arxiv.org/pdf/2002.02492.pdf

Consistency of a Recurrent Language Model

- Completions appear to be *infinite-length*

Prefix *... starboard engines and was going to crash . " We 're going in ,"*

Completion he said . " We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to ...



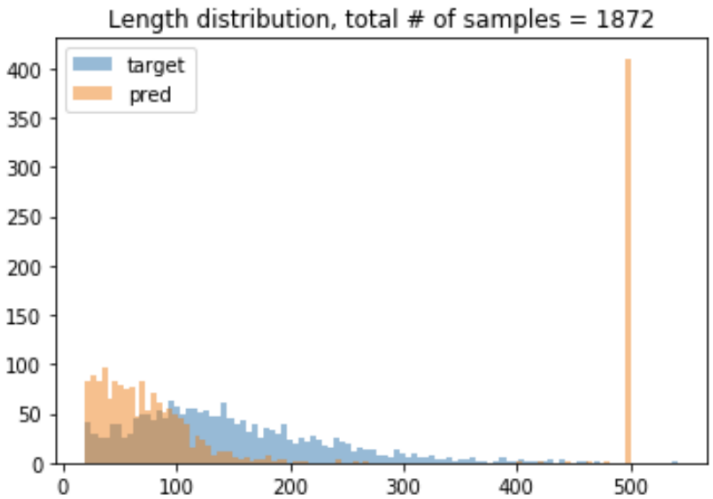
GPT-2 Completions (Wikitext-103 fine-tuned)
https://www.teepublic.com/t-shirt/496713-infinite-symbol

Consistency of a Recurrent Language Model

- Completions appear to be *infinite-length*

Prefix *... starboard engines and was going to crash . " We 're going in ,"*

Completion he said . " We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to ...



- Can we *formally describe* and *analyze* what's happening?
- Can we fix it?

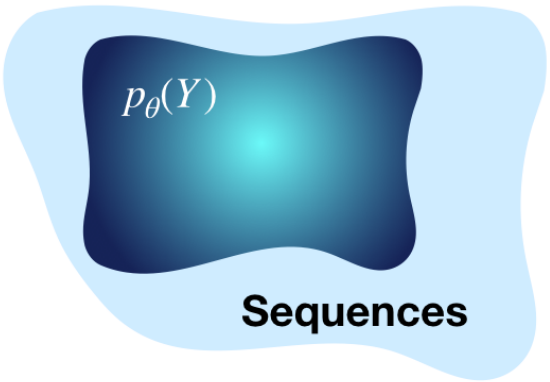
GPT-2 Completions (Wikitext-103 fine-tuned)
https://www.teepublic.com/t-shirt/496713-infinite-symbol

Recurrent Language Model

Definition 2.3 (Recurrent language model). *A recurrent language model p_θ is a neural network that computes the following conditional probability at each time step*

$$p_\theta(y_t = v \mid y_{<t}, C) = \frac{\exp(u_v^\top h_t + c_v)}{\sum_{v' \in V} \exp(u_{v'}^\top h_t + c_{v'})},$$

$$p_\theta(Y \mid C) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, C),$$



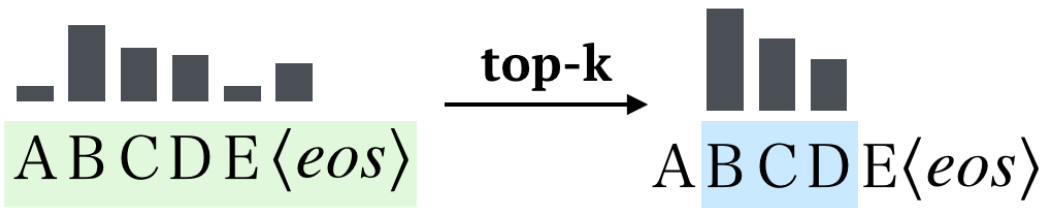
dog whose name is spot . | i have a pet

context

- **Incomplete Decoding Algorithm**
  Only considers a subset of tokens at each time step

Definition 2.11 (Incomplete Decoding). *A decoding algorithm $\mathcal{F}$ is incomplete when for each context C and prefix $y_{<t}$, there is a strict subset $V'_t \subsetneq V$ such that*

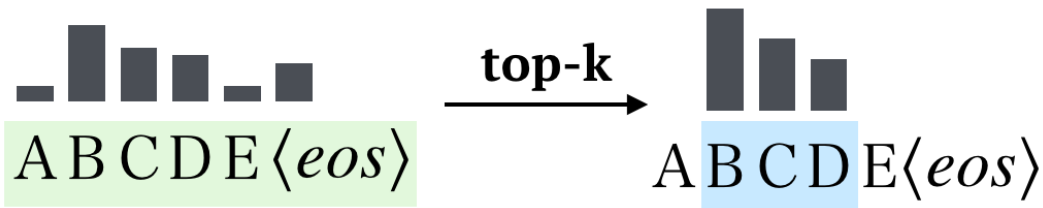
$$\sum_{v \in V'_t} q_{\mathcal{F}}(y_t = v \mid y_{<t}, C) = 1.$$



- **Incomplete Decoding Algorithm**
  Only considers a subset of tokens at each time step

Definition 2.11 (Incomplete Decoding). *A decoding algorithm $\mathcal{F}$ is incomplete when for each context C and prefix $y_{<t}$, there is a strict subset $V_t' \subsetneq V$ such that*

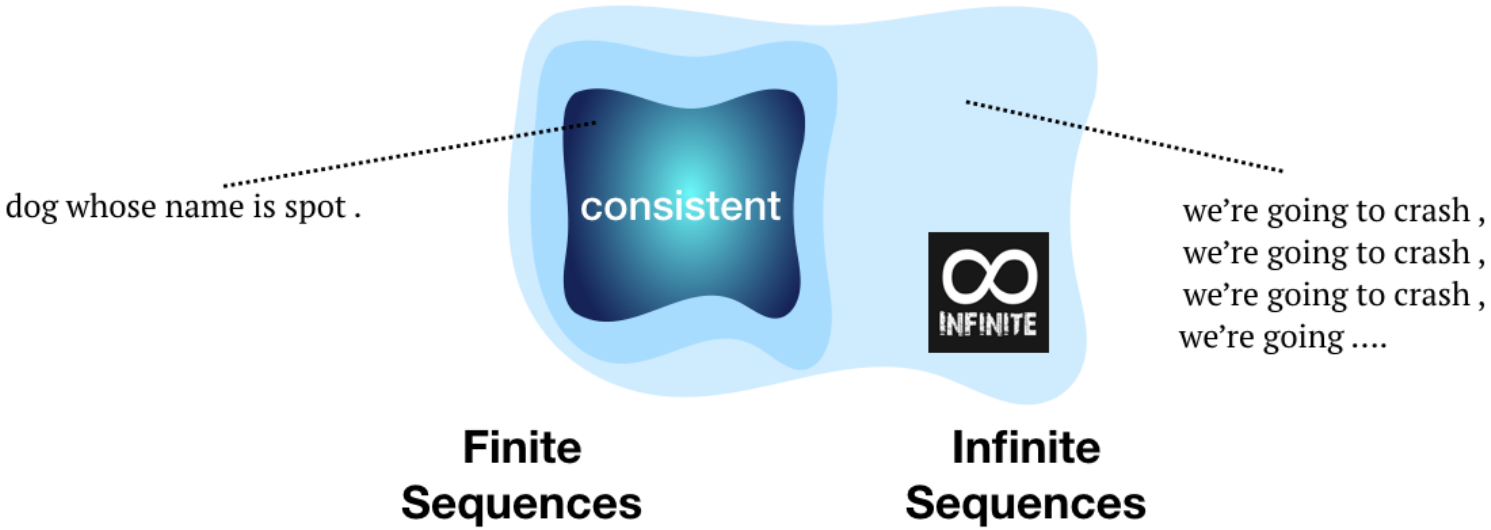
$$\sum_{v \in V_t'} q_{\mathcal{F}}(y_t = v \mid y_{<t}, C) = 1.$$



- Greedy, beam, top-k, nucleus,...

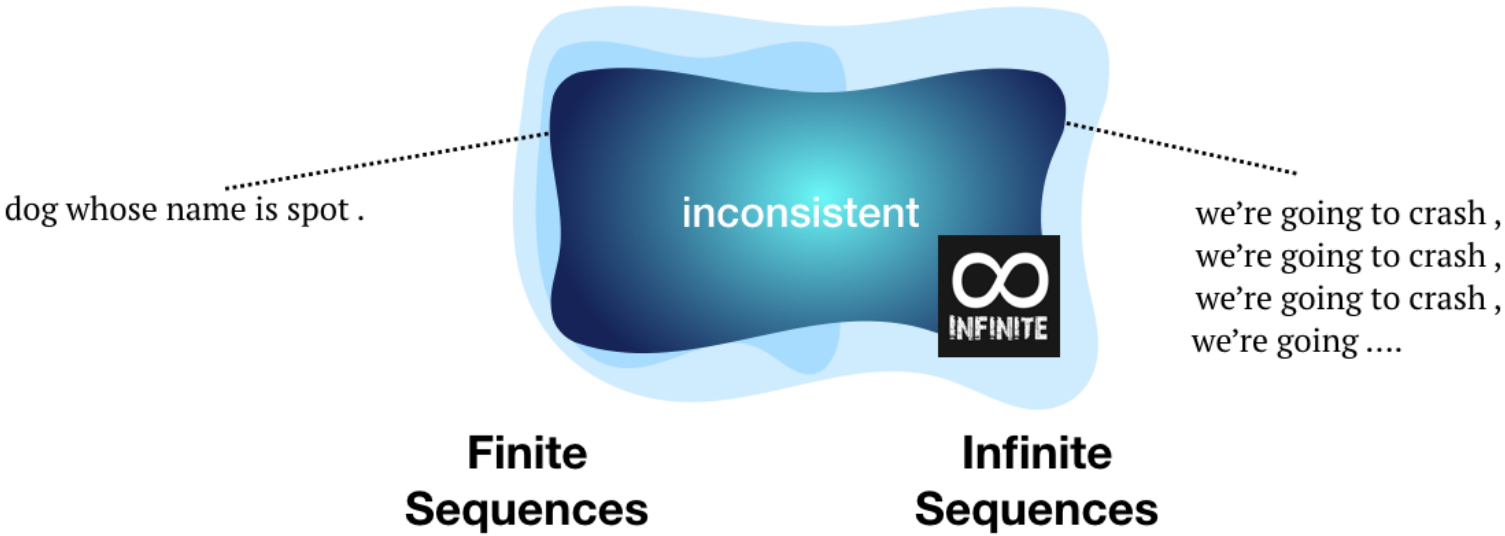
Consistency

Definition 3.1 (Consistency of a recurrent language model). *A recurrent language model is consistent under a context distribution $p(C)$ if $p_\theta(|Y| = \infty) = 0$. Otherwise, the recurrent language model is said to be inconsistent.*



Consistency

Definition 3.1 (Consistency of a recurrent language model)**.** *A recurrent language model is consistent under a context distribution $p(C)$ if $p_\theta(|Y| = \infty) = 0$. Otherwise, the recurrent language model is said to be inconsistent.*



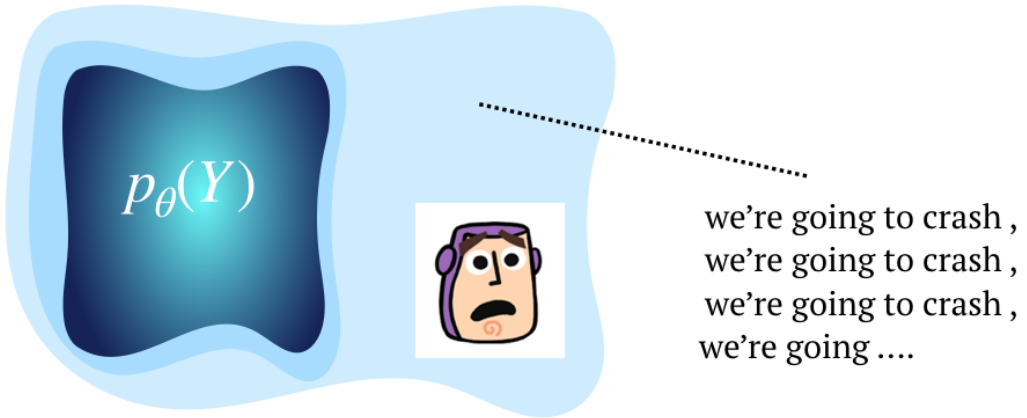
- Is the **model** consistent?[1]

[1] See also [Chen et al 2017, Recurrent Neural Networks as Weighted Language Recognizers]

- Is the **model** consistent?[1] **Yes!**

Lemma 3.2. *A recurrent language model p_θ is consistent if $\|h_t\|_p$ is uniformly bounded for some $p \geq 1$.*

Tanh, layer normalization, ...

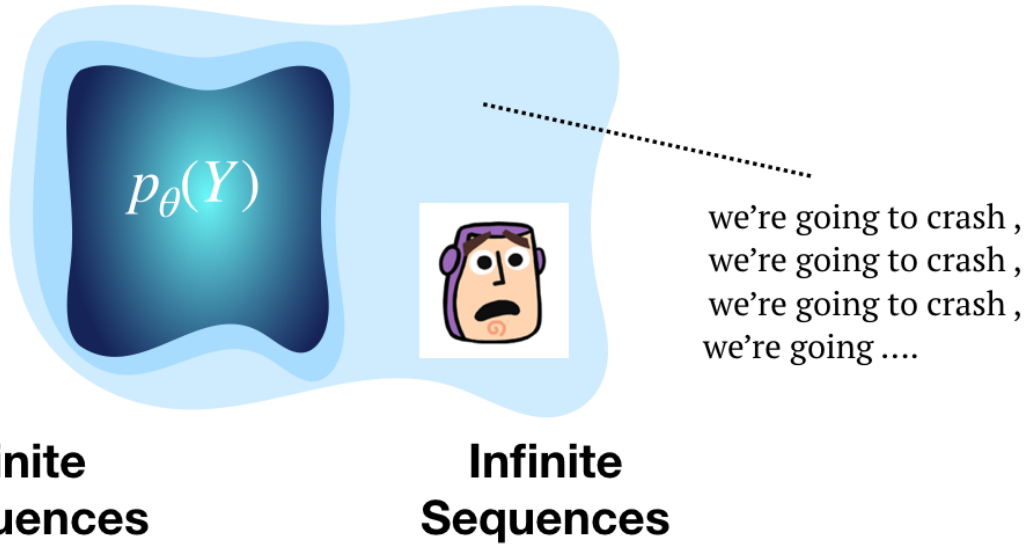


[1] See also [Chen et al 2017, Recurrent Neural Networks as Weighted Language Recognizers]

- Is the **model** consistent?[1] **Yes!**

Lemma 3.2. *A recurrent language model p_θ is consistent if $\|h_t\|_p$ is uniformly bounded for some $p \geq 1$.*

Tanh, layer normalization, ...



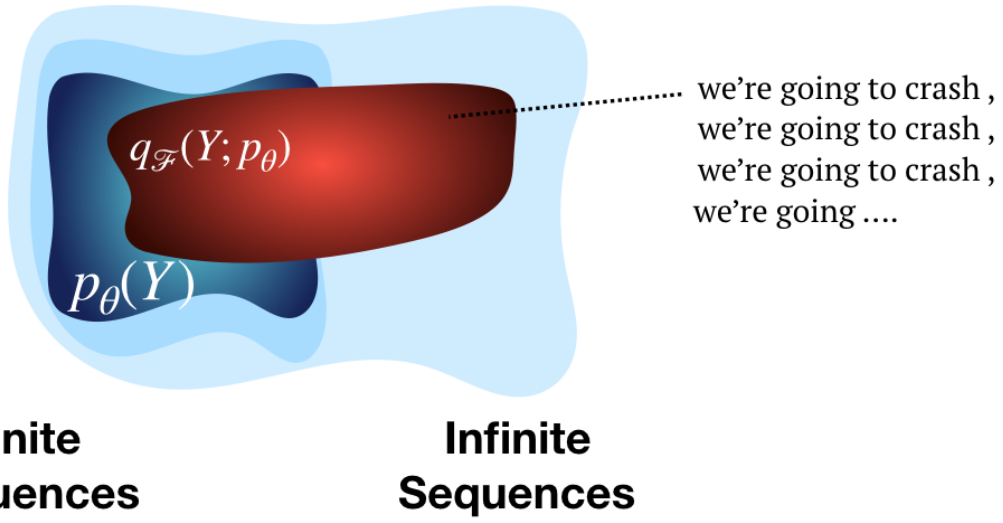
- $p_\theta(\text{infinite sequence}) = 0$

[1] See also [Chen et al 2017, Recurrent Neural Networks as Weighted Language Recognizers]

- Is the **model + decoding algorithm** consistent?

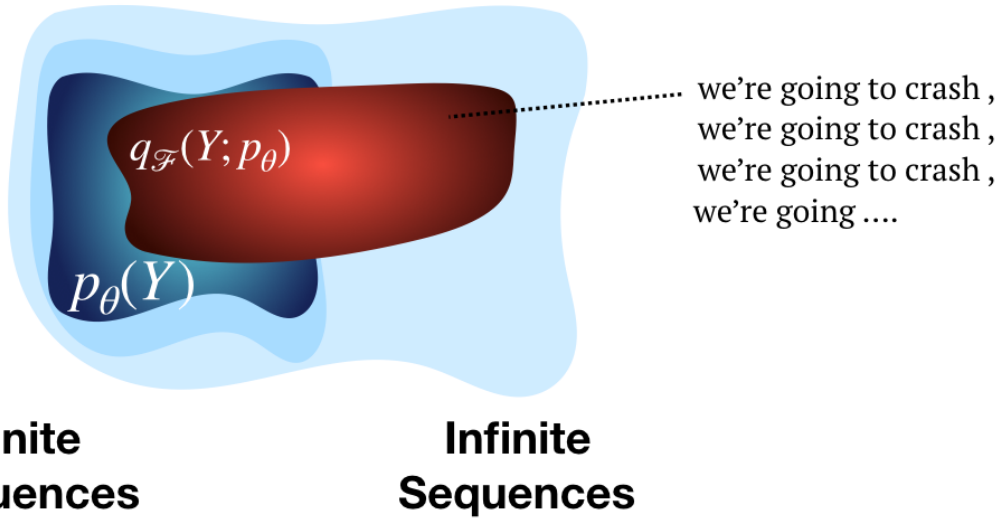
- Is the **model + decoding algorithm** consistent? **No!**

Theorem 3.4 (Inconsistency of an incomplete decoding algorithm). *There exists a consistent recurrent language model p_θ from which an incomplete decoding algorithm $\mathcal{F}$, that considers only up to $(|V| - 1)$-most likely tokens according to $p_\theta(y_t \mid y_{<t}, C)$ at each step t, finds a sequence $\tilde{Y}$ whose probability under p_θ is 0 for any context distribution.*



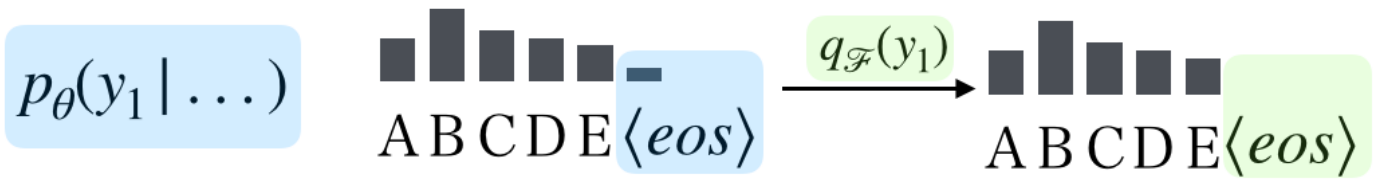
- Is the **model + decoding algorithm** consistent? **No!**

Theorem 3.4 (Inconsistency of an incomplete decoding algorithm). *There exists a consistent recurrent language model p_θ from which an incomplete decoding algorithm $\mathcal{F}$, that considers only up to $(|V| - 1)$-most likely tokens according to $p_\theta(y_t \mid y_{<t}, C)$ at each step t, finds a sequence $\tilde{Y}$ whose probability under p_θ is 0 for any context distribution.*



- Decoded sequences can be infinite-length!

- Is the **model + decoding algorithm** consistent? **No!**
- **Key Idea:** Model might always rank $\langle eos \rangle$ last



RNN always ranks <eos> last .... **<eos> discarded**

.... INFINITE

- In summary, we see **infinite sequences** due to a **mismatch** between the **model** p_θ and **decoded** $q_\mathcal{F}$ distributions.

- In summary, we see **infinite sequences** due to a **mismatch** between the **model** p_θ and **decoded** $q_{\mathcal{F}}$ distributions.
- How do we fix it?

- Ensure $\langle eos \rangle$ is eventually sampled

Consistent top-k sampling

Top-k sampling with:

$$q(v) \propto \begin{cases} p_\theta(v|y_{<t}, C), & \text{if } v \in V', \\ 0, & \text{otherwise,} \end{cases}$$

where $V' = \underbrace{\{\langle eos \rangle\}} \cup \arg\operatorname*{top\text{-}k}_{v'} p_\theta(v' \mid y_{<t}, C)$.

Idea:$\langle eos \rangle$ is always eligible for selection

Consistent top-k sampling

Top-k sampling with:

$$q(v) \propto \begin{cases} p_\theta(v|y_{<t}, C), & \text{if } v \in V', \\ 0, & \text{otherwise,} \end{cases}$$

where $V' = \underbrace{\{\langle eos \rangle\}} \cup \arg\operatorname*{top\text{-}k}_{v'} p_\theta(v' \mid y_{<t}, C)$.

Consistent nucleus sampling

Nucleus sampling with:

$$q(v) \propto \begin{cases} p_\theta(v \mid y_{<t}, C), & \text{if } v \in V_\mu \cup \underbrace{\{\langle eos \rangle\}}, \\ 0, & \text{otherwise.} \end{cases}$$

Idea:$\langle eos \rangle$ is always eligible for selection

- Ensure $\langle eos \rangle$ is eventually top-ranked

Self-terminating Recurrent Language Model

Computes the following conditional probability at each time step:

$$p_\theta(v \mid y_{<t}, C) = \begin{cases} 1 - \alpha(h_t), & \text{if } v = \langle \text{eos} \rangle, \\ \frac{\alpha(h_t) \exp(u_v^\top h_t + c_v)}{\sum_{v' \in V'} \exp(u_{v'}^\top h_t + c_{v'})}, & \text{otherwise,} \end{cases}$$

where

$$\alpha(h_0) = \sigma(u_{\langle \text{eos} \rangle}^\top h_0 + c_{\langle \text{eos} \rangle}),$$
$$\alpha(h_t) = \sigma(u_{\langle \text{eos} \rangle}^\top h_t + c_{\langle \text{eos} \rangle}) \left[1 - p_\theta(\langle \text{eos} \rangle \mid y_{<t-1}, C)\right],$$

with $\sigma : \mathbb{R} \to [0, 1 - \epsilon]$ and $\epsilon \in (0, 1)$.

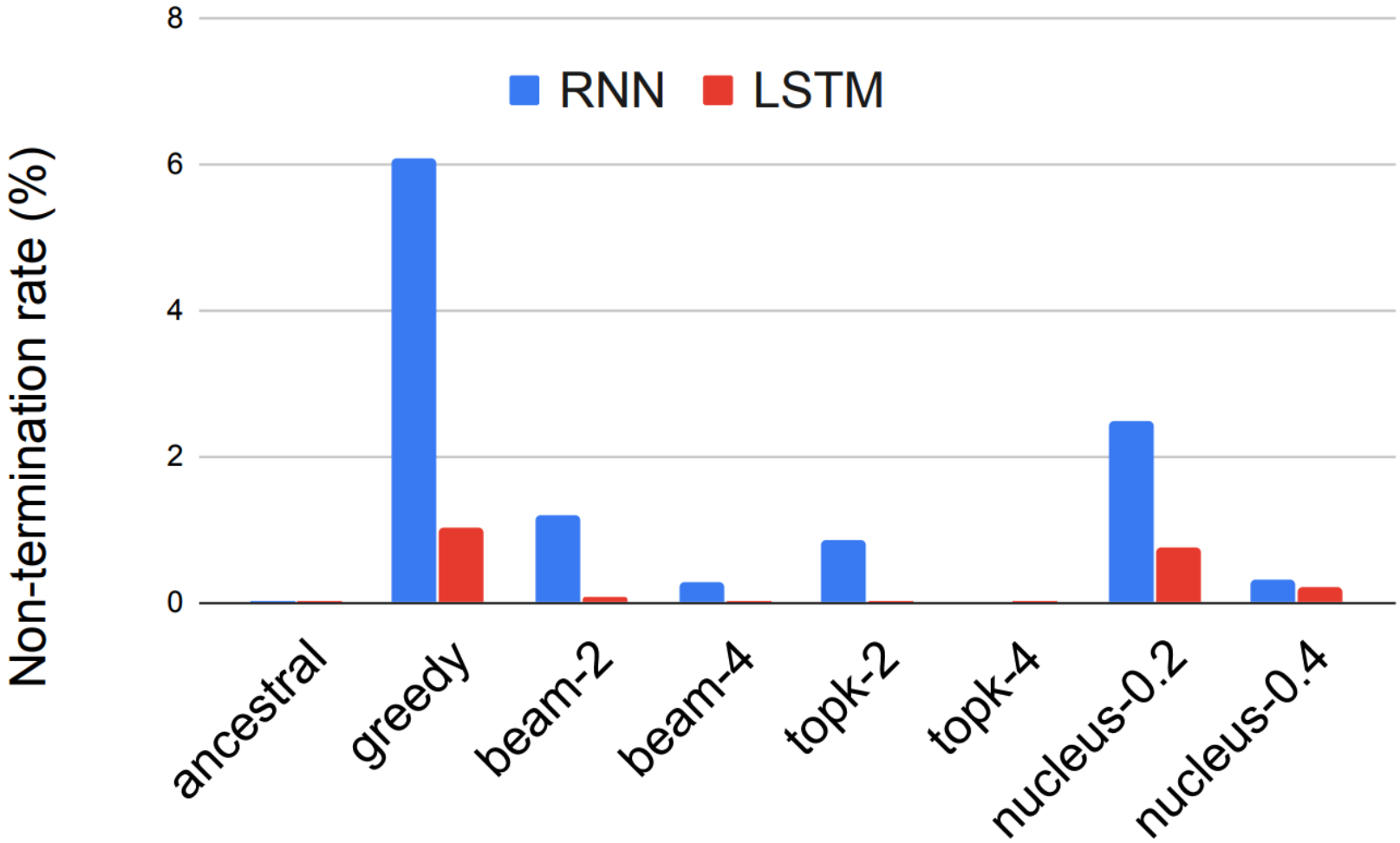
Self-terminating Recurrent Language Model

$$p_t^{\langle \text{eos} \rangle} = 1 - \underbrace{\prod_{t'=0}^{t} \sigma(a_{t'}^{\langle \text{eos} \rangle})}_{\to 0}$$

$p_\theta(\langle eos \rangle | x_{<t}, C)$ increases monotonically.

Does inconsistency occur in practice?

- Wikitext2; non-termination means no ⟨eos⟩ within 1500 steps



Effect of consistent sampling

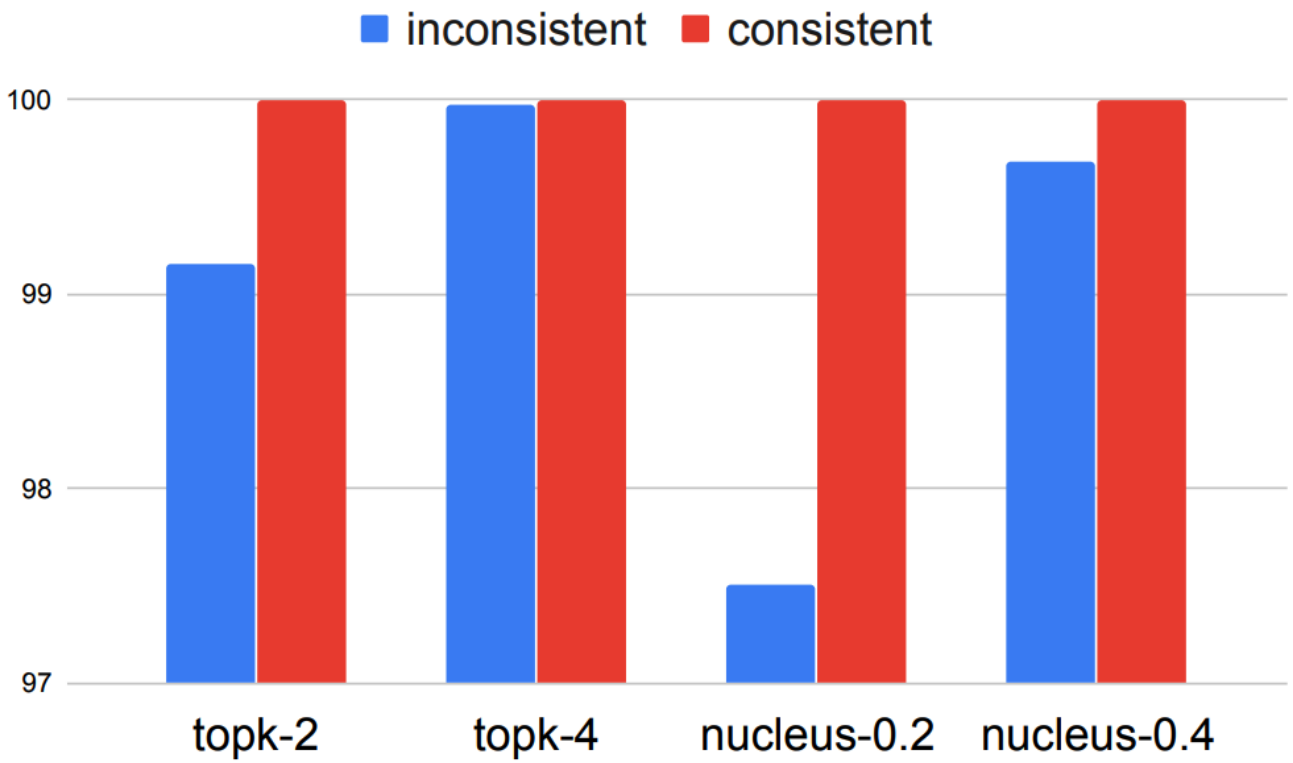


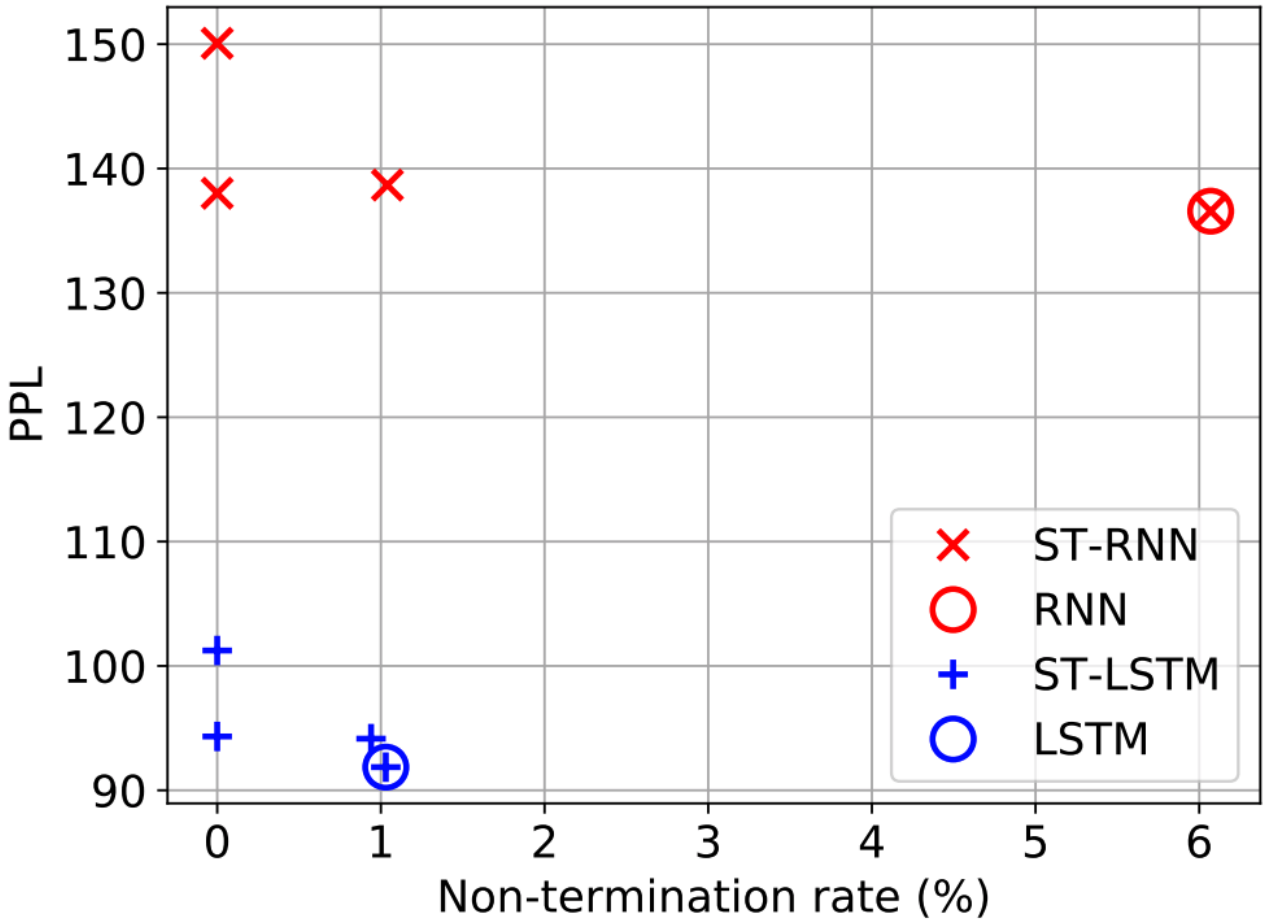
Figure: Termination rate (%) with tanh-RNN

Effect of consistent sampling

Table: Nucleus and consistent nucleus ($\mu = 0.4$) sampling with the LSTM-RNN.

Prefix	*He had a guest @-@ starring role on the television*
nucleus	film the website , with whom he wrote to the title of The Englishwoman 's Domestic Magazine . ⟨eos⟩
c-nucleus	film the website , but he did not be a new sequel . ⟨eos⟩
Prefix	*Somewhere between 29 and 67 species are recognised in the*
nucleus	⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ → ∞
c-nucleus	⟨unk⟩ , with the exception of an average of 6 @.@ 4 in (1 @.@ 6 mm) . ⟨eos⟩
Prefix	*The Civil War saw more ironclads built by both sides*
nucleus	and towns , including ⟨unk⟩ , the British Empire , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ → ∞
c-nucleus	and towns , including ⟨unk⟩ , the British Empire , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ , ⟨unk⟩ ⟨eos⟩

Effect of Self-terminating Recurrent LM



Effect of Self-terminating Recurrent LM

Table: LSTM-RNN and a self-terminating LSTM-RNN ($\epsilon = 10^{-3}$)

Prefix	*With 2 : 45 to go in the game ,*
Baseline	the team was able to gain a first down . ⟨eos⟩
STRLM	the Wolfpack was unable to gain a first down . ⟨eos⟩
Prefix	*As of 2012 , she is a horse riding teacher*
Baseline	, and a ⟨unk⟩ , and a ⟨unk⟩ , and a ⟨unk⟩ , and a ⟨unk⟩ , and a ⟨unk⟩ , and a ⟨unk⟩ , and a ⟨unk⟩ , and a ⟨unk⟩ , and a ⟨unk⟩ , $\rightarrow \infty$
STRLM	, and a member of the ⟨unk⟩ ⟨unk⟩ . ⟨eos⟩
Prefix	*Nintendo Power said they enjoyed Block Ball and its number*
Baseline	of songs , including the " ⟨unk⟩ " , " ⟨unk⟩ " , " ⟨unk⟩ " , . . . , " Ode to a Nightingale " , " Ode to a Nightingale ", " Ode to a $\rightarrow \infty$
STRLM	of songs , including the " ⟨unk⟩ " , " ⟨unk⟩ " , . . . {" ⟨unk⟩ " ,}^{45} . . . " ⟨unk⟩ " , " ⟨eos⟩

Recap

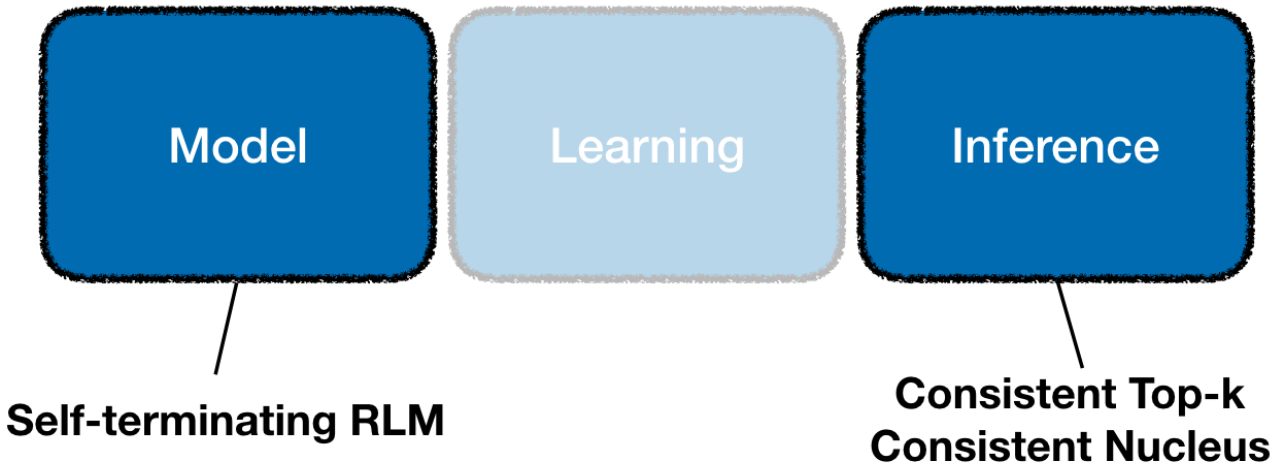
- Text degeneration is characterized by **non-terminating sequences**, **repetition**, and **logical contradictions**

Recap

- Text degeneration is characterized by **non-terminating sequences**, **repetition**, and **logical contradictions**
- Common models + decoding algorithms can yield **infinite sequences** due to ranking $\langle \text{eos} \rangle$ low and excluding it

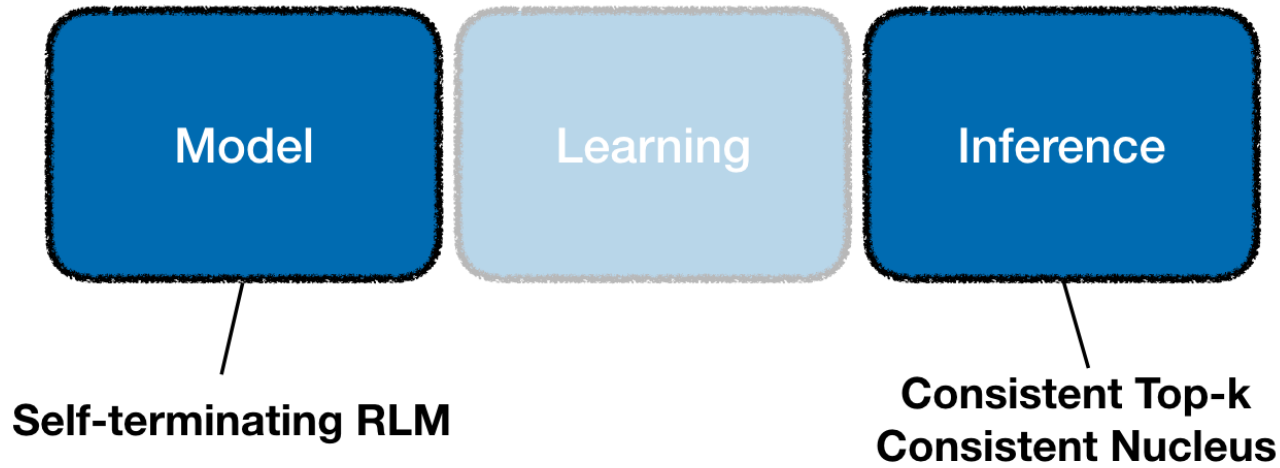
Recap

- Text degeneration is characterized by **non-terminating sequences**, **repetition**, and **logical contradictions**
- Common models + decoding algorithms can yield **infinite sequences** due to ranking ⟨eos⟩ low and excluding it
- New **model** and **inference** methods can address the issue



Recap

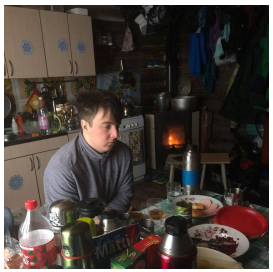
- Text degeneration is characterized by **non-terminating sequences**, **repetition**, and **logical contradictions**
- Common models + decoding algorithms can yield **infinite sequences** due to ranking $\langle \text{eos} \rangle$ low and excluding it
- New **model** and **inference** methods can address the issue



Next: **Repetition & Learning**

Neural Text Generation with Unlikelihood Training

Welleck, Kulikov, Roller, Dinan, Cho, Weston ICLR 2020

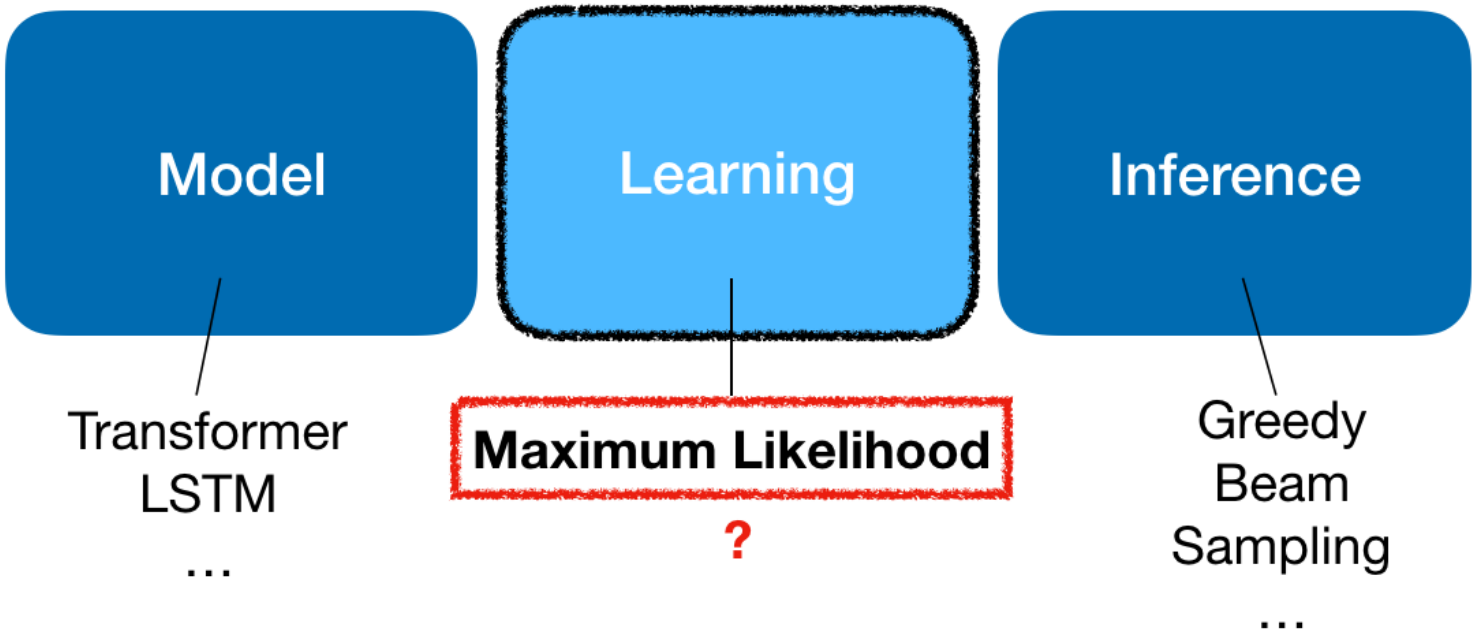




https://arxiv.org/pdf/1908.04319.pdf

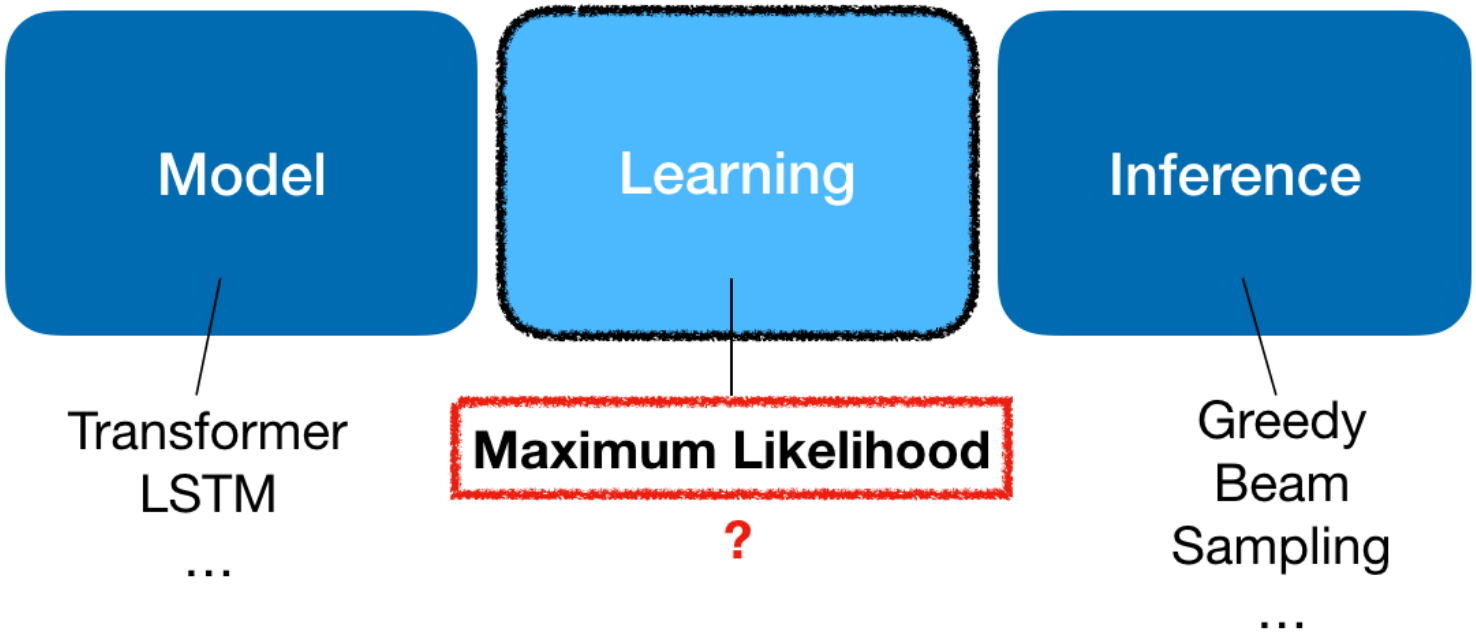
Neural Text Generation with Unlikelihood Training

[Welleck, Kulikov et al ICLR 2020]



Neural Text Generation with Unlikelihood Training

[Welleck, Kulikov et al ICLR 2020]



- Is there a learning objective that doesn't lead to degeneration?
  - 1st step: for **specified** degenerate behavior (e.g. repetition).

Maximum Likelihood Estimation (MLE)

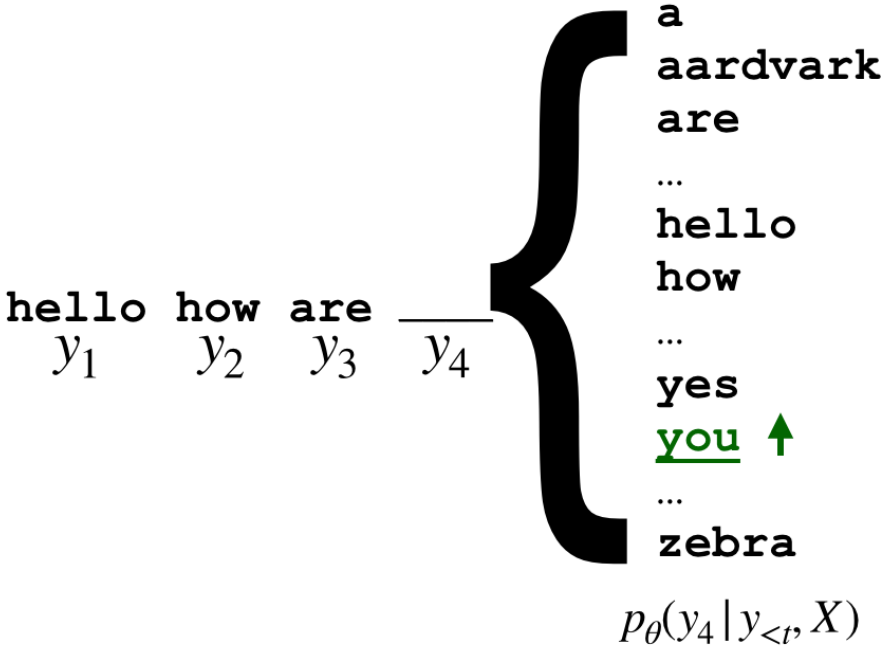
Given training example (X, Y),

$$\mathcal{L}_{\text{MLE}} = -\sum_{t=1}^{|Y|} \log p_\theta(y_t | y_{<t})$$

Maximum Likelihood Estimation (MLE)

Given training example (X, Y),

$$\mathcal{L}_{\text{MLE}} = -\sum_{t=1}^{|Y|} \log p_\theta(y_t | y_{<t})$$



Maximum Likelihood Estimation (MLE)

Prefix	*... starboard engines and was going to crash . " We 're going in ,"*
$\mathcal{L}_{\text{MLE}}$	he said . " We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to

Repetition

Model: 40% 4-gram repetition

Human: ~1% 4-gram repetition

Large-Scale Transformer LM [from Welleck et al ICLR 2020]

Unlikelihood (Token-level)

$$\mathcal{L}_{\text{ULE}} = \mathcal{L}_{\text{MLE}} + \alpha \mathcal{L}_{\text{UL}}$$

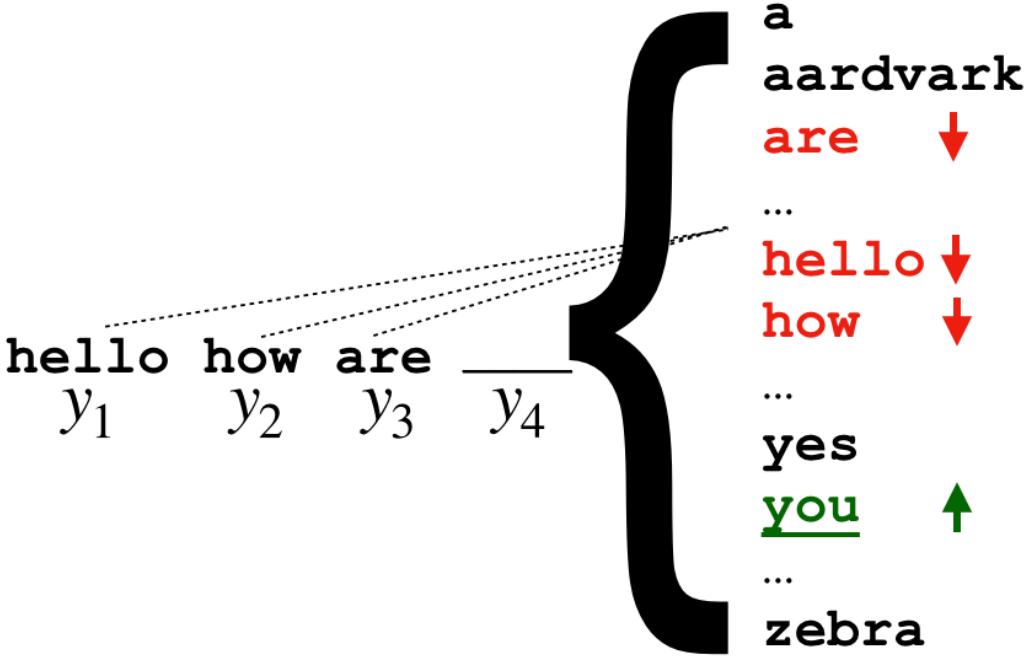
$$\mathcal{L}_{\text{UL}}^{t} = - \sum_{y_{\text{neg}} \in \mathcal{C}_t} \log \left(1 - p_\theta(y_{\text{neg}} | y_{<t}, X)\right)$$

Unlikelihood (Token-level)

$$\mathcal{L}_{\text{ULE}} = \mathcal{L}_{\text{MLE}} + \alpha \mathcal{L}_{\text{UL}}$$

$$\mathcal{L}_{\text{UL}}^{t} = - \sum_{y_{\text{neg}} \in \mathcal{C}_t} \log \left(1 - p_\theta(y_{\text{neg}} | y_{<t}, X)\right)$$

$$\mathcal{C}_t = \{y_1, \ldots, y_{t-1}\} \backslash \{y_t\}$$



Unlikelihood (Sequence-level)

$$\underbrace{\hat{Y}}_{\text{decoded sequence}} \sim \mathcal{F}(p_\theta, X)$$

$$\mathcal{L}_{\text{ULE}} = \mathcal{L}_{\text{MLE}}(X, Y) + \alpha \mathcal{L}_{\text{UL}}(X, \hat{Y}, \mathcal{C})$$

Unlikelihood (Sequence-level)

$$\underbrace{\hat{Y}}_{\text{decoded sequence}} \sim \mathcal{F}(p_\theta, X)$$

$$\mathcal{L}_{\text{ULE}} = \mathcal{L}_{\text{MLE}}(X, Y) + \alpha \mathcal{L}_{\text{UL}}(X, \hat{Y}, \mathcal{C})$$

Candidates $\mathcal{C}$:

- Random $\quad \mathcal{C}_t = \{\hat{y}_t\}$ if *rand*() else $\emptyset$

he said . we're going to crash . we're going to crash

$\hat{Y}$

Unlikelihood (Sequence-level)

$$\underbrace{\hat{Y}}_{\text{decoded sequence}} \sim \mathcal{F}(p_\theta, X)$$

$$\mathcal{L}_{\text{ULE}} = \mathcal{L}_{\text{MLE}}(X, Y) + \alpha \mathcal{L}_{\text{UL}}(X, \hat{Y}, \mathcal{C})$$

Candidates $\mathcal{C}$:

- Repeats $\mathcal{C}_t = \{\hat{y}_t\}$ if $\hat{y}_t \in$ repeat, else $\emptyset$

`he said . we're going to crash . we're going to crash`

$\hat{Y}$

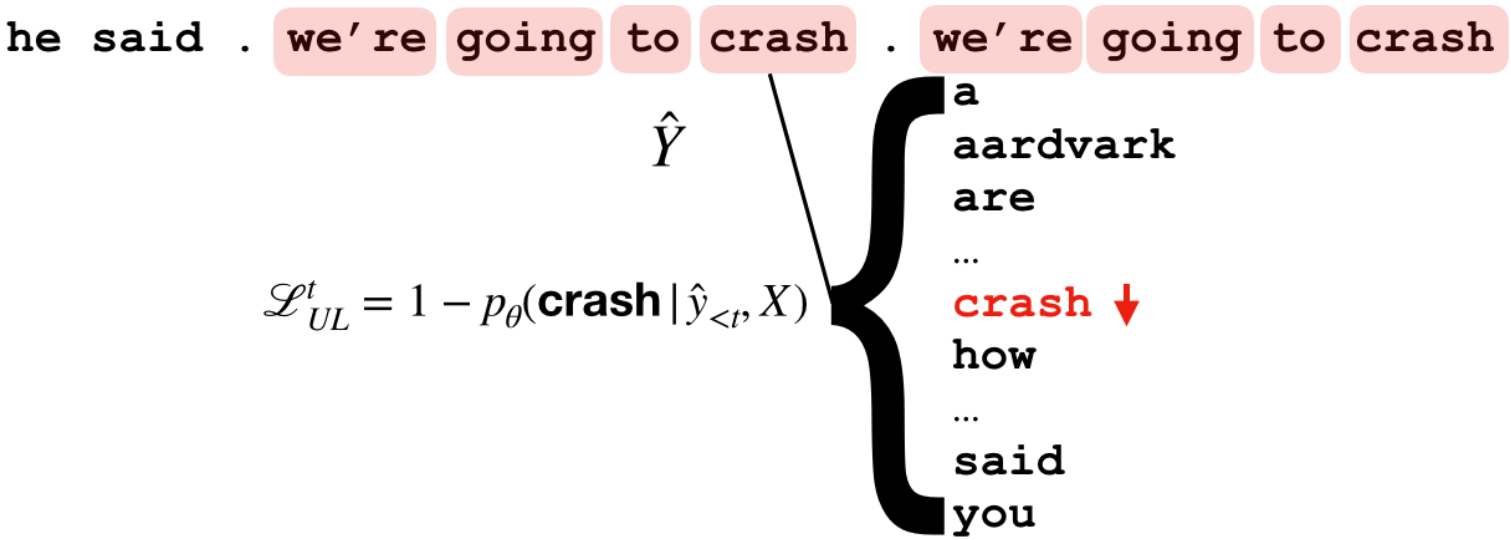
Unlikelihood (Sequence-level)

$$\underbrace{\hat{Y}}_{\text{decoded sequence}} \sim \mathcal{F}(p_\theta, X)$$

$$\mathcal{L}_{\text{ULE}} = \mathcal{L}_{\text{MLE}}(X, Y) + \alpha \mathcal{L}_{\text{UL}}(X, \hat{Y}, \mathcal{C})$$

Candidates $\mathcal{C}$:

- Repeats $\mathcal{C}_t = \{\hat{y}_t\}$ if $\hat{y}_t \in$ repeat, else $\emptyset$



		4-gram repetition
Prefix	*... Lyrically the song has excerpts of different languages including French , Spanish*	
MLE	, Italian , Spanish , Italian , Spanish , Italian , Spanish , Spanish , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Spanish , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese	0.744
Unlikelihood	, and German . In the first verse , the protagonist sings about being a " girl who 's been in love with someone else " , while the second verse describes the relationship between the protagonist and her lover . In the third verse , the protagonist sings	0.063
Prefix	*... starboard engines and was going to crash . " We 're going in ,"*	
MLE	he said . " We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to	0.787
Unlikelihood	Hood said . " I 'm going to make sure we 're going to get back to the water . " The order to abandon ship was given by Admiral Beatty , who ordered the remaining two battlecruisers to turn away . At 18 : 25 , Hood turned his	0.000

Model	search	seq-rep-4	uniq-seq	ppl
MLE	greedy	.442	10.8k	**25.64**
	beam	.523	9.5k	
UL - Random	greedy	.116	12.5k	25.518
	beam	.146	14.2k	
UL - Repeat	greedy	**.058**	**15.4k**	26.72
	beam	**.013**	**19.1k**	
Human	-	.006	19.8k	-

Winner		Loser	Crowdworkers Win rate	Experts Win rate
Unlikelihood	*beats*	**MLE**	*82%	
Unlikelihood	*beats*	**MLE + Nucleus Sampling (0.9)**	59%	*83%
Unlikelihood	*beats*	**MLE + Beam Blocking (4 gram)**	60%	*74%

Why does it work?

$$\nabla \mathcal{L}_a = x^* - m \odot p, \quad m_i = \begin{cases} (1 - \alpha \frac{p_{\text{neg}}}{1 - p_{\text{neg}}}) & \text{if } i \neq i_{\text{neg}} \\ (1 + \alpha) & \text{if } i = i_{\text{neg}}, \end{cases}$$

1. **Negative candidate**: negative gradient
2. **Ground-truth token**: positive gradient
3. **Others**: positive when negative candidate probability is high

Why does it work?

- Noise contrastive estimation [Gutmann & Hyvrinen 2010]

$$\underbrace{\mathbb{E}_{x_+ \sim p_+}\left[\log h_\theta(x_+)\right]}_{\textit{positives}} + \underbrace{\mathbb{E}_{x_- \sim p_-}\left[\log(1 - h_\theta(x_-))\right]}_{\textit{negatives}}$$

Where

$$h_\theta(x_-) = \sigma(\log p_\theta(x) - \log p_-(x))$$

Why does it work?

- Noise contrastive estimation [Gutmann & Hyvrinen 2010]

$$\underbrace{\mathbb{E}_{x_+ \sim p_+}\left[\log h_\theta(x_+)\right]}_{\textit{positives}} + \underbrace{\mathbb{E}_{x_- \sim p_-}\left[\log(1 - h_\theta(x_-))\right]}_{\textit{negatives}}$$

Where

$$h_\theta(x_-) = \sigma(\log p_\theta(x) - \log p_-(x))$$

- Unlikelihood with random candidates:

$$\underbrace{\mathbb{E}_{\substack{x \sim p_+ \\ t \sim \mathcal{U}(T)}}\left[\log p_\theta(x_t|x_{<t})\right]}_{\textit{positives}} + \underbrace{\mathbb{E}_{\substack{x \sim p_- \\ t \sim \mathcal{U}(T)}}\left[\log\left(1 - p_\theta(x_t|x_{<t})\right)\right]}_{\textit{negatives}}.$$

Follow-up: Unlikelihood for Dialogue

Don't Say *That*!
Making Inconsistent Dialogue Unlikely with Unlikelihood Training

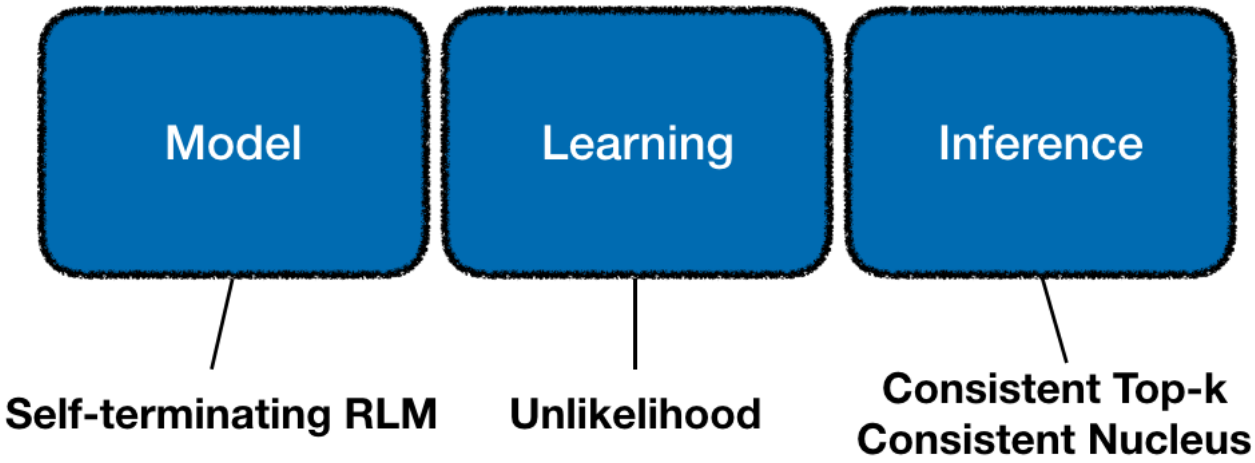
Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck
Y-Lan Boureau, Kyunghyun Cho, Jason Weston
Facebook AI Research

- **Repetition *and* logical contradictions**

https://arxiv.org/pdf/1911.03860.pdf

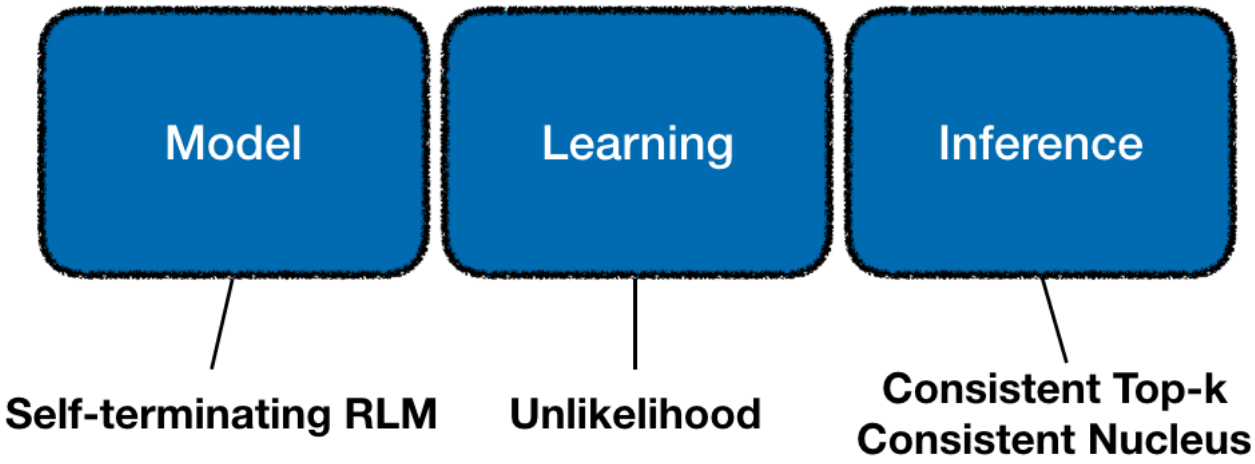
Recap

- Text degeneration is characterized by **non-terminating sequences**, **repetition**, and **logical contradictions**
- New **model** and **inference** methods can address the issues



Recap

- Text degeneration is characterized by **non-terminating sequences**, **repetition**, and **logical contradictions**
- New **model** and **inference** methods can address the issues



But more investigation and methods are needed!

Thank you
very much very much very much very much very $\rightarrow \infty$